Transformers and multilinguality

Course outline

1. Motivation for Transformer. Attention. The original Transformer architecture.
2. Transformer-based Encoders. Masked language models based on the Transformer architecture. BERT and related models.
3. Classification and sequence tagging with Transformers. Using encoders to generate feature representation for various NLU tasks.
4. Transformer-based Decoders. Generation of text based on the Transformer architecture. GPT and related decoders. Text generation methods. Prompt tuning.
5. Prompt and Instruction tuning. Reinforcement Learning from Human Feedback (RLHF), ChatGPT, and related models.
6. Sequence to sequence tasks: machine translation, text detoxification, question answering, dialogue. Technical tricks for training and inference: infrastructure and performance.
7. Multilingual language models based on the Transformer architecture.
8. Efficient Transformers.
9. Compression of transformer models.
10. Network encoders with Transformers.
11. Multimodal and vision Transformers.
12. Transformers for tabular data.
13. Transformers for event sequences.

Agenda

- Multilingual resources, tasks and difficulties
- Multilingual models
- Tricks and recipes for multilingual NLP

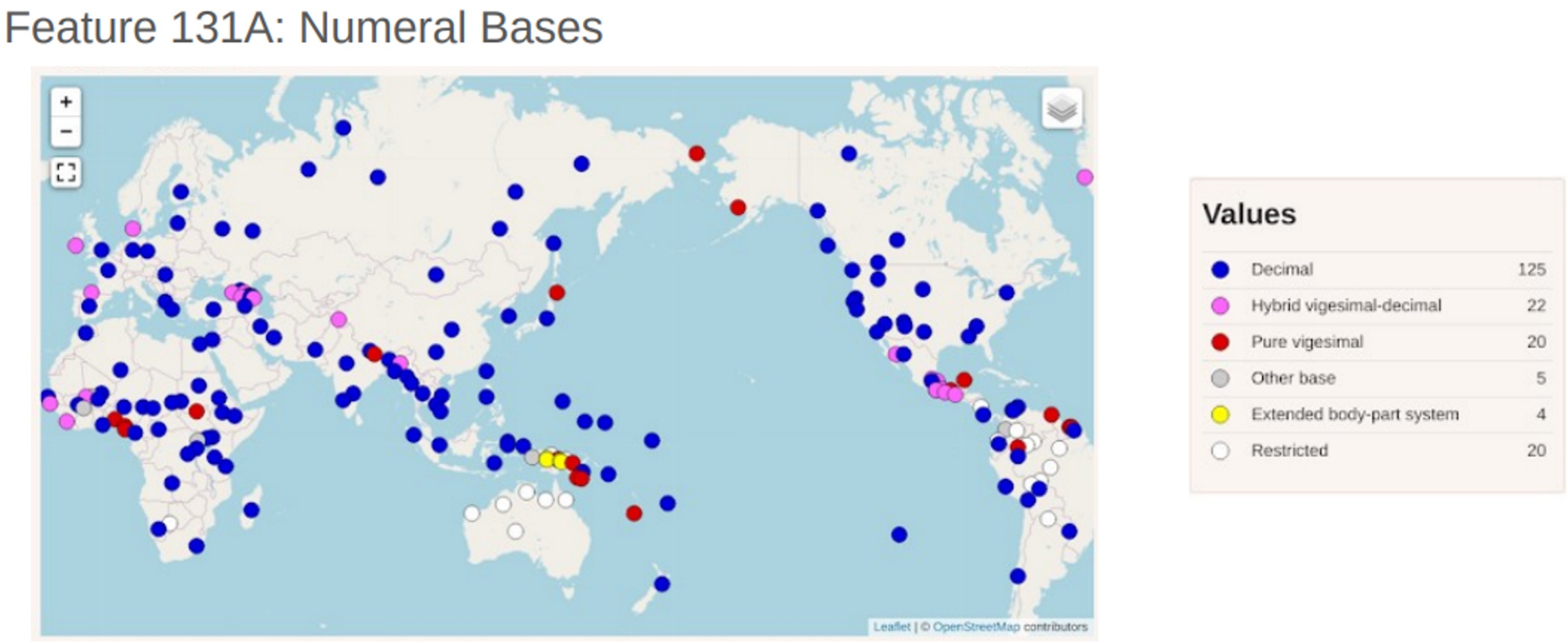
Partially based on CMU CS 11-737 Multilingual NLP course http://phontron.com/class/multiling2022

Why multilingual?

- Languages other than English or Russian do exist.
  - And as empires fall apart, new languages get official status and wider usage
  - Users want their content in their own languages
- It is expensive to support separate NLP models for each language
- Most languages are “low-resource”
  - Monolingual models for them are often not good enough
  - But we can transfer NLP knowledge across languages
    - For closely related languages (e.g. ru->by), it can be transferred directly
    - For more distant languages, translation might be required
      - For good translation, we need parallel corpora to train
      - To collect parallel corpora, we need good NLU models…

The language space: WALS typology

The World Atlas of Language Structures stores unified language *features*



http://phontron.com/class/multiling2022/assets/slides/multiling-03-typology.pdf ; https://wals.info/

The language space: WALS typology

- 2,676 languages, 192 attributes

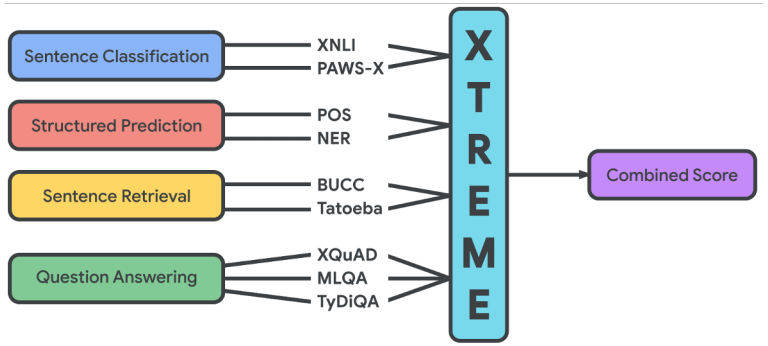
ID#	Feature Name	Category	Feature Values
1	Consonant Inventories	Phonology (19)	{1:Large, 2:Small, 3:Moderately Small, 4:Moderately Large, 5:Average}
23	Locus of Marking in the Clause	Morphology (10)	{1:Head, 2:None, 3:Dependent, 4:Double, 5:Other}
30	Number of Genders	Nominal Categories (28)	{1:Three, 2:None, 3:Two, 4:Four, 5:Five or More}
58	Obligatory Possessive Inflection	Nominal Syntax (7)	{1:Absent, 2:Exists}
66	The Perfect	Verbal Categories (16)	{1:None, 2:Other, 3:From 'finish' or 'already', 4:From Possessive}
81	Order of Subject, Object and Verb	Word Order (17)	{1:SVO, 2:SOV, 3:No Dominant Order, 4:VSO, 5:VOS, 6:OVS, 7:OSV}
121	Comparative Constructions	Simple Clauses (24)	{1:Conjoined, 2:Locational, 3:Particle, 4:Exceed}
125	Purpose Clauses	Complex Sentences (7)	{1:Balanced/deranked, 2:Deranked, 3:Balanced}
138	Tea	Lexicon (10)	{1:Other, 2:Derived from Sinitic 'cha', 3:Derived from Chinese 'te'}
140	Question Particles in Sign Languages	Sign Languages (2)	{1:None, 2:One, 3:More than one}
142	Para-Linguistic Usages of Clicks	Other (2)	{1:Logical meanings, 2:Affective meanings, 3:Other or none}

Example from Georgi, Xia and Lewis (2010)

http://phontron.com/class/multiling2022/assets/slides/multiling-03-typology.pdf ; https://wals.info/

Examples of multilingual tasks

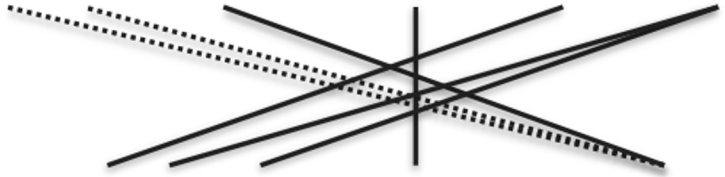
- Translation between multiple languages (e.g. FLORES)
- MASSIVE NLU benchmark in 51 language from Amazon Alexa
  - Recognize intents and slots in dialogues with assistant in any language
- NeuCLIR benchmark in cross-language information retrieval
  - Search among Zh, Fa and Ru documents with En queries
- Multilingual News Article Similarity
- Multilingual Complex Named Entity Recognition
- Composite benchmarks: XTREME, XGLUE



Why is it difficult to translate?

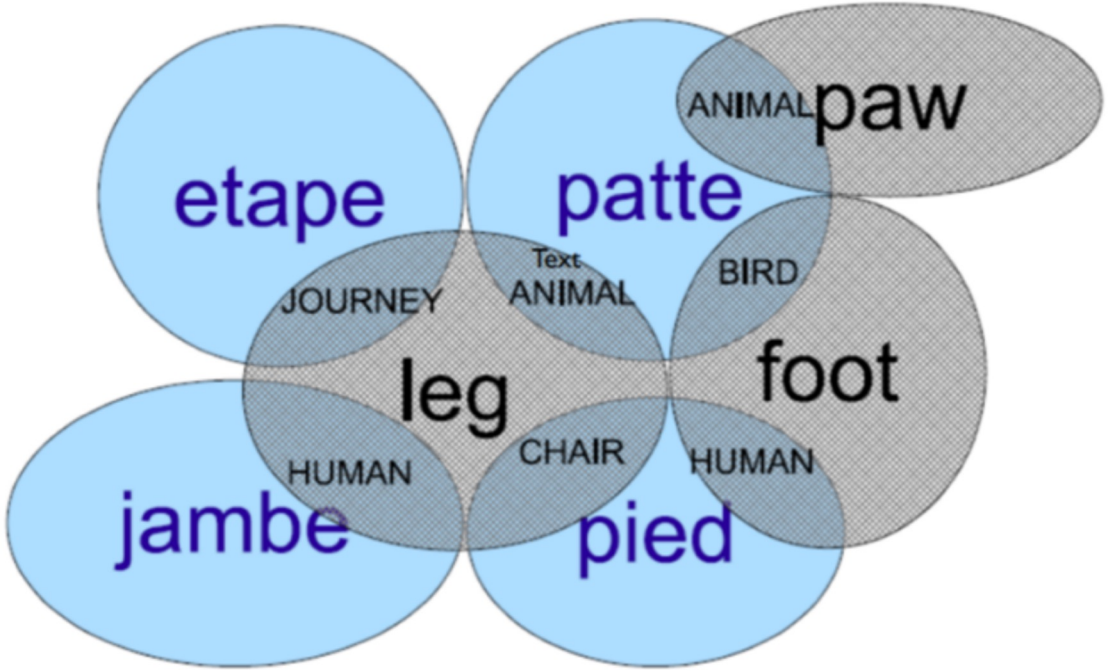
in the in-city exploded a car-bomb

German: In der Innenstadt explodierte eine Autobombe



English: A car bomb exploded downtown.

Translationese: In the inner city, there exploded a car bomb.



ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

Эти типы стали есть на складе

- Материал находится на складе
- Люди едят на складе
- Сталь нужно есть на складе

Images source: https://web.stanford.edu/~jurafsky/slp3/

Examples of parallel corpora

- Important books
  - Bible, Tanzil (Quran)
- Governmental texts
  - Europarl, UN corpus, etc.
- Subtitles
  - OpenSubtitles, TED, etc.
- Computer manuals
  - PHP, Ubuntu, etc.
- Aligned web data
  - ParaCrawl, WikiMatrix, CCMatrix, etc.
- A major repository: OPUS

Multilingual models

How multilingual is multilingual BERT?

- The most typical multilingual pretrained models are BERT-like
  - E.g. multilingual BERT (2018), XLM(2019), XLM-R (2020), mDeBERTaV3 (2021)
- Most of them (except XLM) are fully unsupervised
- Still, they can perform cross-language transfer
- How does it even work???
  - Common vocabulary
  - Some mapping between vocabularies of similar languages
    - E.g. Hindi (Devanagari script) vs Urdu (Arabic script)
  - Generalization depends on the number of shared WALS features
- Perhaps, alignment occurs via shared words (e.g. URLs)

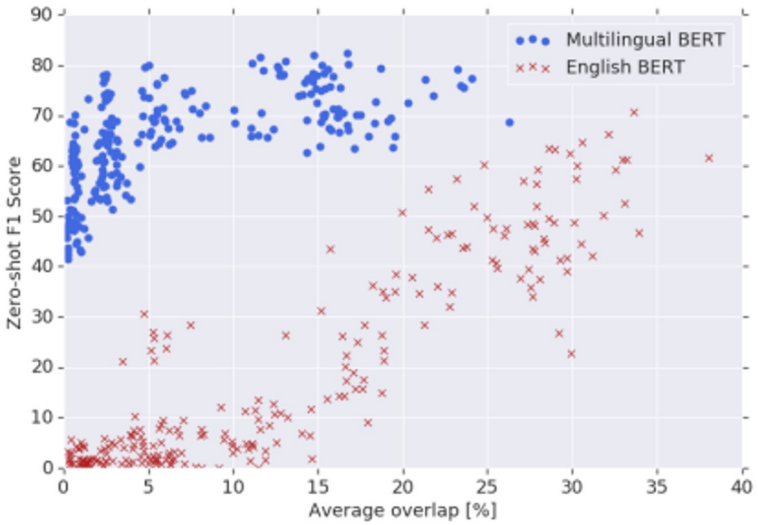


Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

How multilingual is multilingual BERT? Pires et al, 2019, https://arxiv.org/abs/1906.01502

Multilingual generation

- XGLM by Meta
  - Pretrain a GPT-like model on a balanced corpus of 30 languages
  - Probe with few-shot in-context learning
    - SOTA in some generation tasks for lower-resourced languages
    - Capable of few-shot translation
- mGPT by Sber
  - Pretrained on 61 languages from Wikipedia and MC4
  - High scores in many zero-shot and few-shot tasks
- Both models can be fine-tuned for specific tasks

https://arxiv.org/abs/2112.10668 ; https://arxiv.org/abs/2204.07580

Multilingual seq2seq models

- mT5
  - Pretrained with a standard monolingual T5 denoising objective on mC4 (100 langs)
  - Fine-tuned for each task separately
  - SOTA on some multilingual NLI, NER and QA benchmarks
- mBART
  - Pretrained with a standard BART denoising objective on 25 languages; later extended to 50 languages
  - Language id is specified by the BOS token
  - Fine-tuned on translation pairs, achieved SOTA on low- and mid-resource languages
- M2M100 and related models
  - An mBART-like transformer trained to translate between 2200 language pairs and 100 languages
  - The encoder produces nearly language-agnostic embeddings

https://arxiv.org/abs/2010.11934 ; https://arxiv.org/abs/2001.08210; https://arxiv.org/abs/2010.11125

Adapting BERTs to new languages

- Simplest: fine-tune the model on the target language
- Vocabulary adaptation: more efficient, but more complex
  - Remove unused tokens from the vocabulary (based on a target-lang corpus)
  - Add new tokens (e.g. by adding producing some BPE merges)
  - Initialize new embeddings using average embeddings of their constituents or source-language tokens aligned with them
  - Fine-tune the model on the target-lang (with e.g. MLM loss)
    - To speed it up, only embeddings can be fine-tuned (at least, for the 1st epoch)
- Training from scratch (which is more expensive)

https://aclanthology.org/W19-3712; https://arxiv.org/abs/1905.07213; https://aclanthology.org/2022.acl-short.68/

Tips for multilingual classification

- Augmentation with translated data helps
- Domain and task adaptation usually helps
- Multilingual training usually helps
- Zero-shot transfer works OK, but worse than

Model	Data	DE	FR	JA	ES
multi-target	target	94.1	93.8	91.1	78.1
multi-all	all	93.8	94.3	91.4	77.7
zero-shot	EN	92.7	92.6	88.5	72.1

			CLS					HATEVAL				
Model	**Adapt.**	**Aug.**	**EN**	**DE**	**FR**	**JA**	**AVG**	**EN**	**EN†**	**ES**	**AVG**	**AVG†**
mono-target												
RoBERTa (EN) BERT (OTHERS)	×	×	$94.7_{0.4}$	$90.9_{0.6}$	$95.2_{0.0}$	$88.7_{0.3}$	92.4	$44.4_{5.3}$	$58.5_{6.2}$	$75.6_{0.6}$	60.0	67.1
		✓	$\mathbf{95.3}_{0.3}$	$92.0_{0.2}$	$95.6_{0.3}$	$89.3_{0.02}$	93.0	$46.1_{2.6}$	$60.6_{3.2}$	$76.0_{1.7}$	61.0	68.3
	TAPT	×	$94.9_{0.1}$	$91.6_{0.1}$	$95.4_{0.1}$	$89.3_{0.3}$	92.8	$45.4_{1.9}$	$59.9_{2.7}$	$76.1_{1.1}$	60.8	68.0
		✓	$95.0_{0.4}$	$92.3_{0.4}$	$95.8_{0.2}$	$89.7_{0.4}$	93.2	$44.7_{1.5}$	$59.2_{1.7}$	$76.9_{1.4}$	60.8	68.0
	TAPT+	×	$94.9_{0.4}$	$91.8_{0.2}$	$95.5_{0.3}$	$89.5_{0.2}$	92.9	$48.0_{1.5}$	$63.1_{2.6}$	$76.3_{1.1}$	62.2	69.7
	DAPT	✓	$\mathbf{95.3}_{0.1}$	$93.0_{0.8}$	$\mathbf{95.9}_{0.1}$	$89.9_{0.4}$	**93.5**	$46.0_{4.3}$	$60.2_{4.4}$	$76.9_{0.6}$	61.4	68.5
multi-target												
XLM-RoBERTa	×	×	$92.5_{0.4}$	$93.0_{0.2}$	$92.5_{0.3}$	$90.4_{0.5}$	92.1	$47.2_{2.0}$	$61.4_{1.9}$	$74.8_{0.5}$	61.0	68.1
		✓	$93.3_{0.1}$	$94.0_{0.2}$	$93.8_{0.2}$	$90.3_{0.3}$	92.8	$45.6_{1.6}$	$59.3_{2.5}$	$77.0_{1.1}$	61.3	68.1
	TAPT	×	$92.7_{0.5}$	$93.5_{0.5}$	$93.9_{0.3}$	$90.3_{0.1}$	92.6	$47.0_{2.7}$	$62.4_{3.3}$	$76.1_{1.4}$	61.6	69.2
		✓	$93.4_{0.6}$	$94.0_{0.3}$	$93.8_{0.5}$	$90.5_{0.4}$	92.9	$47.9_{1.3}$	$63.5_{1.5}$	$77.9_{0.9}$	62.9	70.7
	TAPT+	×	$93.1_{0.6}$	$93.0_{0.5}$	$93.6_{0.1}$	$90.8_{0.3}$	92.6	$49.9_{2.5}$	$65.6_{2.4}$	$76.5_{1.0}$	63.2	71.0
	DAPT	✓	$94.0_{0.3}$	$\mathbf{94.1}_{0.4}$	$93.8_{0.3}$	$91.1_{0.4}$	93.2	$46.6_{2.1}$	$61.7_{2.5}$	$\mathbf{78.1}_{0.8}$	62.3	69.9
multi-all												
XLM-RoBERTa	×	×	$92.4_{0.3}$	$92.6_{0.4}$	$93.3_{0.4}$	$90.4_{0.4}$	92.2	$48.4_{3.5}$	$63.1_{4.5}$	$77.5_{0.4}$	62.9	70.3
		✓	$93.4_{0.3}$	$93.3_{0.2}$	$94.0_{0.2}$	$90.4_{0.5}$	92.8	$49.8_{3.5}$	$66.0_{4.6}$	$77.8_{0.9}$	63.8	71.9
	TAPT	×	$92.5_{0.4}$	$93.0_{0.3}$	$93.9_{0.3}$	$90.9_{0.3}$	92.6	$48.4_{2.7}$	$64.2_{3.5}$	$77.4_{0.9}$	62.9	70.8
		✓	$93.5_{0.4}$	$93.4_{0.5}$	$94.1_{0.2}$	$91.1_{0.2}$	93.0	$50.0_{2.2}$	$66.5_{2.6}$	$77.8_{0.6}$	63.9	72.2
	TAPT+	×	$92.7_{0.3}$	$93.3_{0.2}$	$94.0_{0.3}$	$91.2_{0.3}$	92.8	$47.1_{3.9}$	$62.7_{5.3}$	$77.4_{1.0}$	62.3	70.1
	DAPT	✓	$93.5_{0.3}$	$93.8_{0.2}$	$94.3_{0.3}$	$\mathbf{91.4}_{0.2}$	93.3	$\mathbf{50.7}_{1.1}$	$\mathbf{67.4}_{1.4}$	$77.7_{0.7}$	**64.2**	**72.6**

Practical Transformer-based Multilingual Text Classification, Wang and Banco, 2021: https://aclanthology.org/2021.naacl-industry.16

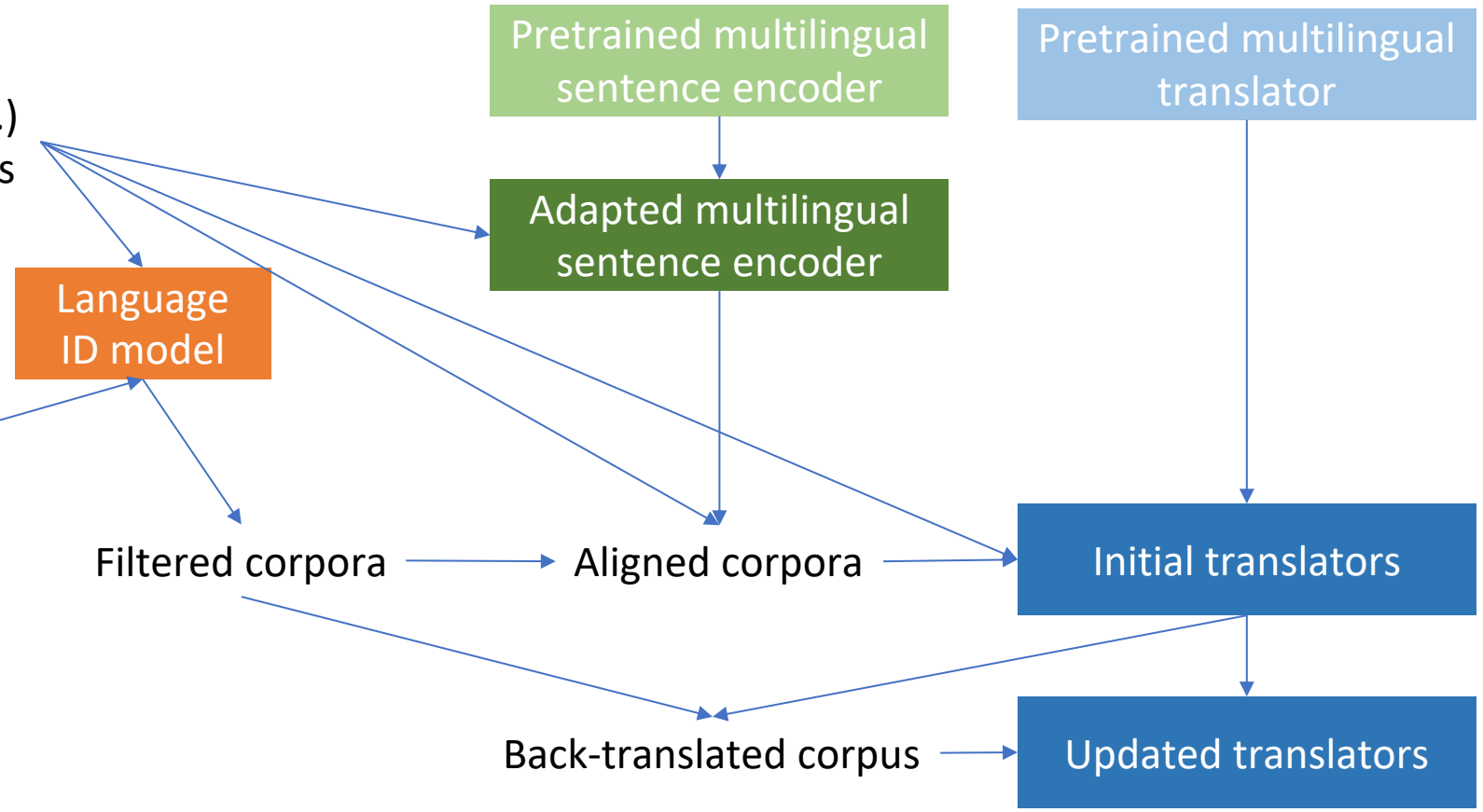
How to bootstrap NLP for a new language?

Typical initial resources:

- Small parallel data (bible, laws, etc.)
- Word- and phrase-level dictionaries
- Wikipedia

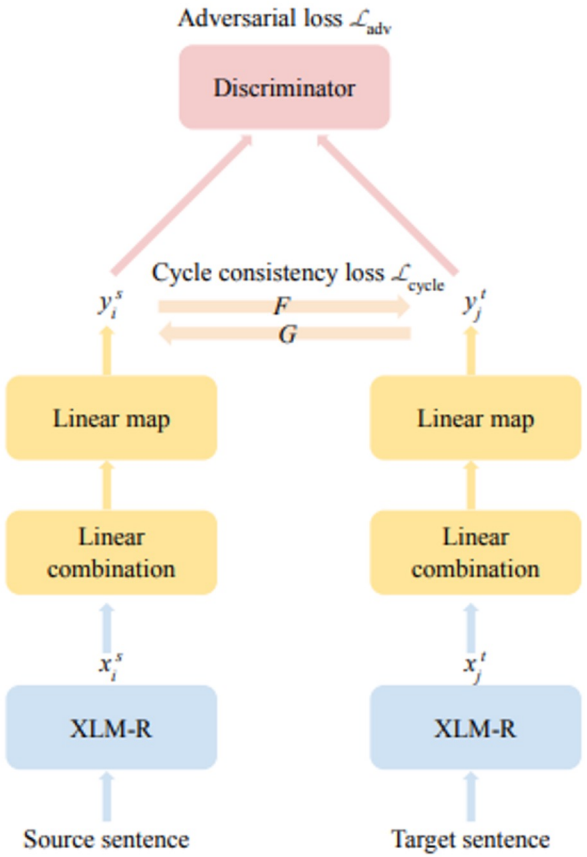
Dirty corpora

- Wikipedia in other languages
- Mixed-language web crawl
- Unaligned parallel literature



XLM-R –> universal sentence encoder

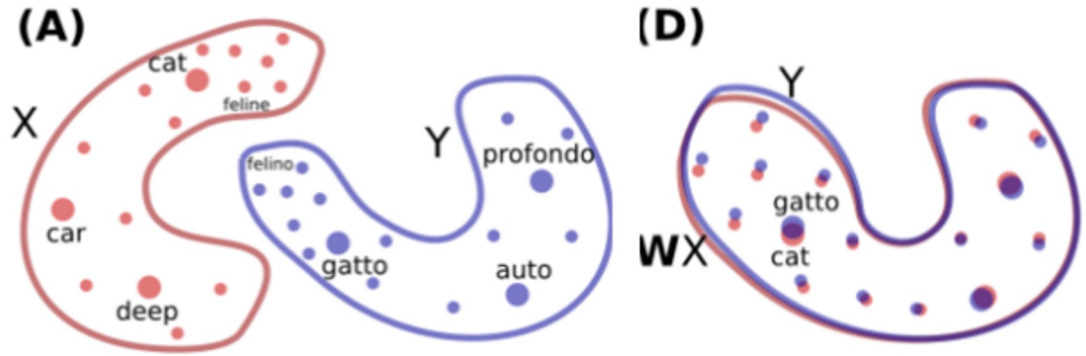
- Use XLM-R as a fixed feature extractor
- Train a weighted average pooler and FFN head to extract cross-lingual embeddings
  - It is possible to train it even an unsupervised way: with cycle consistency and adversarial loss
- Such a model can be used for matching sentences in unseen languages
  - (Because XLM-R has already seen them)



https://arxiv.org/abs/2104.07642

Unsupervised word translation

- Hypothesis: Word embedding spaces in two languages are isomorphic
  - One embedding space can be linearly transformed into another
  - Give monolingual embeddings X and Y, learn a (orthogonal) matrix, such that, WX = Y
- Use adversarial learning to learn W:
  - If WX and Y are perfectly aligned, a discriminator shouldn't be able to tell
  - Discriminator: Predict whether an embedding is from Y or the transformed space WX.
  - Train W to confuse the discriminator



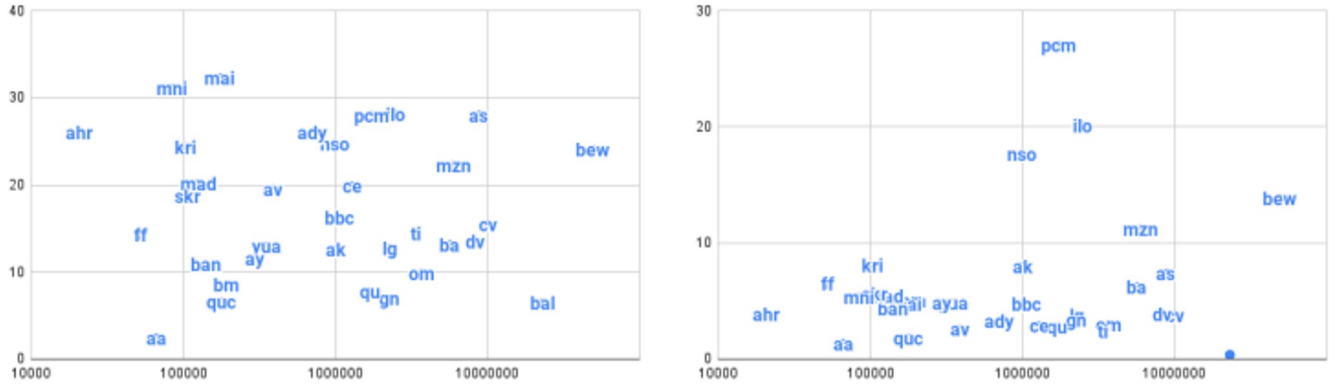
After aligning words, a sentence translation model can be trained:

- Pretrain with monolingual denoising
- Finetune with back-translation

http://phontron.com/class/multiling2022/assets/slides/multiling-10-unsupervisedmt.pdf

What is next?

- *Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning*, a paper by Google
- The bootstrapping pipeline for 1000 languages
  - Language identification
  - Monolingual denoising pretraining in all languages
  - Fine-tuning on en->xx and xx->en pairs
  - Good translation for some zero-resource languages



(a) Any-to-English (xx → en)

(b) English-to-Any (en → xx)

Figure 2: Unsupervised/zero-resource BLEU on 30 new languages. The x-axis depicts the amount of monolingual data available for the language, while the y-axis depicts the BLEU score of the 1.6B parameter Transformer model after fine-tuning with online back-translation. The data point corresponding to each language is represented by its BCP-47 language code.

https://arxiv.org/abs/2201.03110

BigScience Large Open-science Open-access Multilingual Language Model

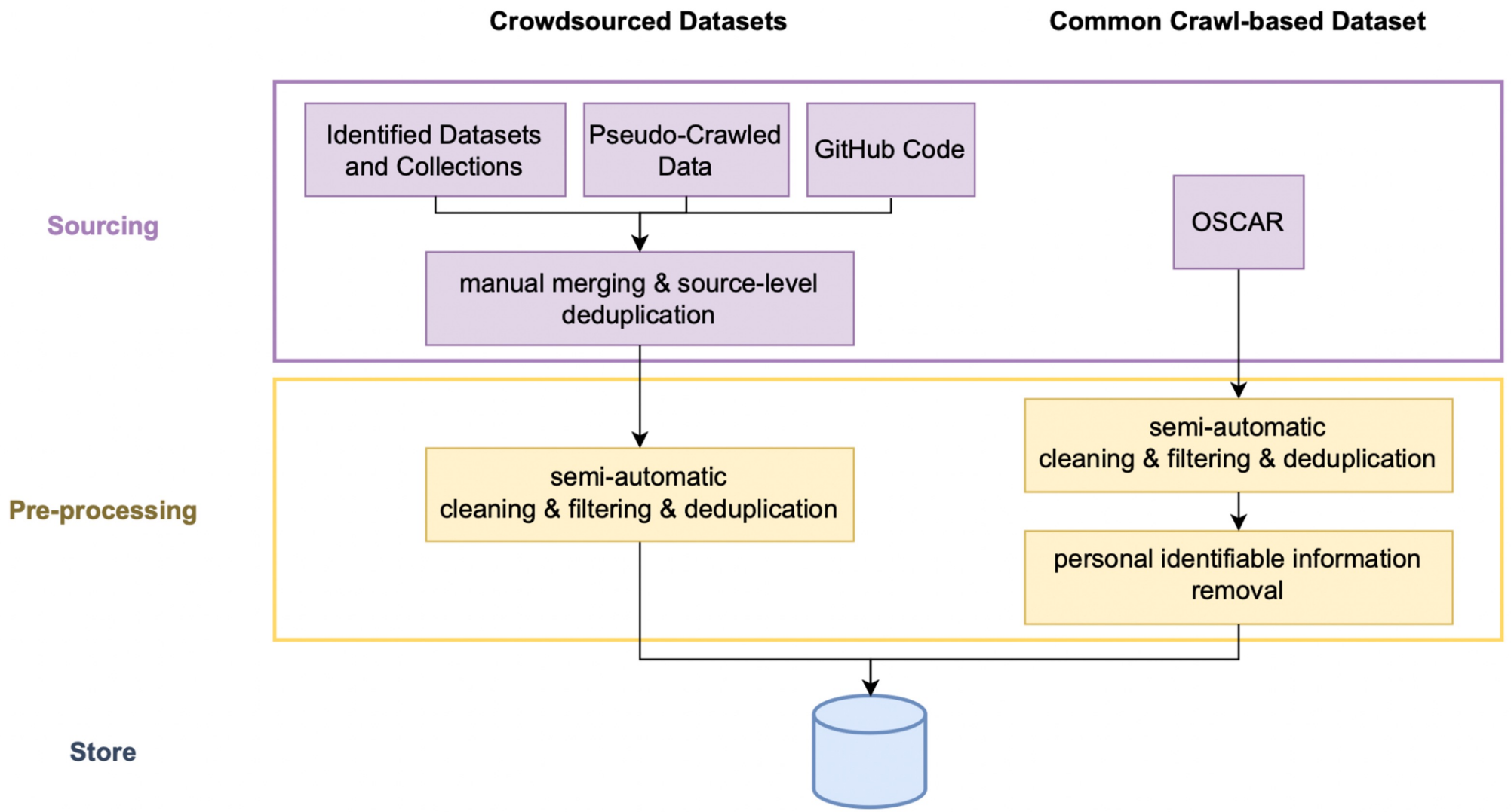
https://arxiv.org/pdf/2211.05100.pdf

BigScience Large Open-science Open-access Multilingual Language Model

- a 176 billion parameter language model;
- trained on 46 natural languages and 13 programming languages;
- final run of 117 days (March 11 - July 6) training
- developed and released by a collaboration of 1000 researchers from 70+ countries and 250+ institutions;
- on the Jean Zay supercomputer in the south of Paris, France;
- compute grant worth an estimated €3M from French research agencies CNRS and GENCI;

https://arxiv.org/pdf/2211.05100.pdf

Dataset



https://arxiv.org/pdf/2211.05100.pdf

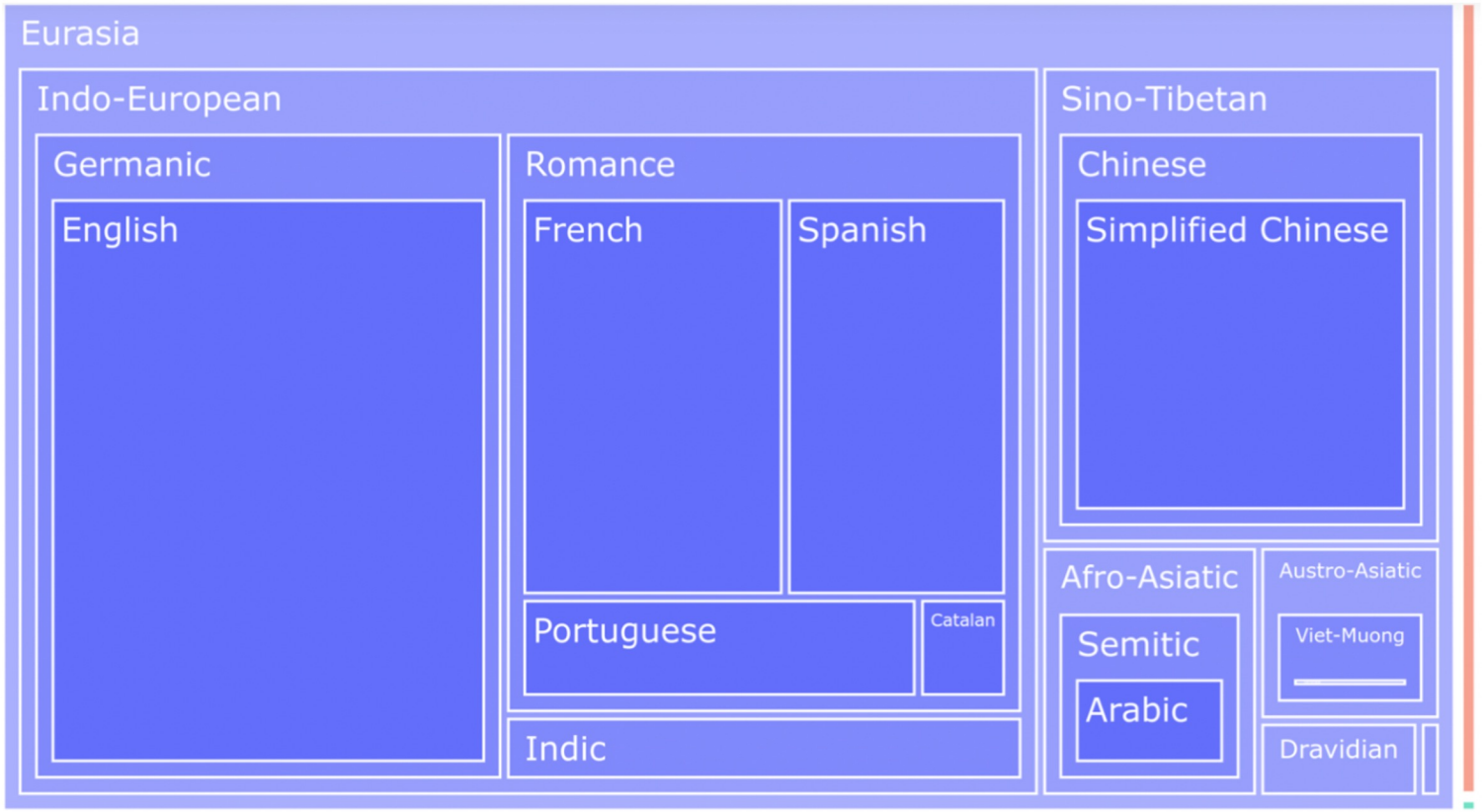
Dataset

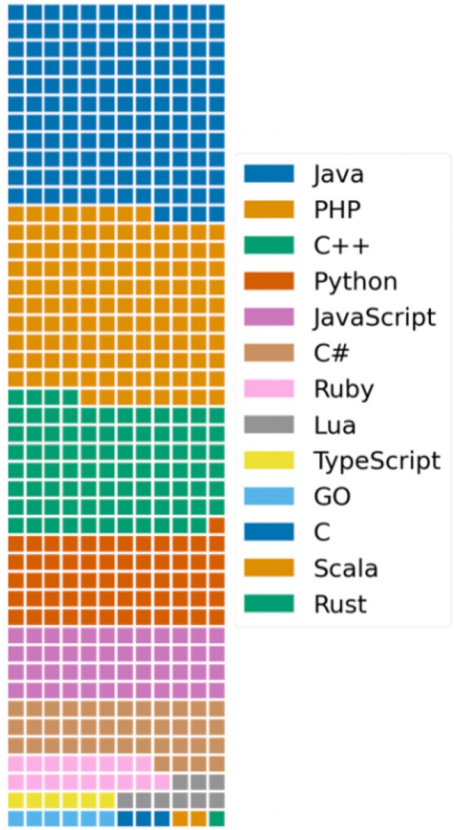
ROOTS corpus (Laurenon et al., 2022)

- a composite collection of 498 Hugging Face datasets
- 1.61 terabytes of text that span 46 natural languages and 13 programming languages.

https://arxiv.org/pdf/2211.05100.pdf

Dataset





https://arxiv.org/pdf/2211.05100.pdf

Language	ISO-639-3	catalog-ref	Genus	Family	Macroarea	Size in Bytes
Akan	aka	ak	Kwa	Niger-Congo	Africa	70,1554
Arabic	arb	ar	Semitic	Afro-Asiatic	Eurasia	74,854,900,600
Assamese	asm	as	Indic	Indo-European	Eurasia	291,522,098
Bambara	bam	bm	Western Mande	Mande	Africa	391,747
Basque	eus	eu	Basque	Basque	Eurasia	2,360,470,848
Bengali	ben	bn	Indic	Indo-European	Eurasia	18,606,823,104
Catalan	cat	ca	Romance	Indo-European	Eurasia	17,792,493,289
Chichewa	nya	ny	Bantoid	Niger-Congo	Africa	1,187,405
chiShona	sna	sn	Bantoid	Niger-Congo	Africa	6,638,639
Chitumbuka	tum	tum	Bantoid	Niger-Congo	Africa	170,360
English	eng	en	Germanic	Indo-European	Eurasia	484,953,009,124
Fon	fon	fon	Kwa	Niger-Congo	Africa	2,478,546
French	fra	fr	Romance	Indo-European	Eurasia	208,242,620,434
Gujarati	guj	gu	Indic	Indo-European	Eurasia	1,199,986,460
Hindi	hin	hi	Indic	Indo-European	Eurasia	24,622,119,985
Igbo	ibo	ig	Igboid	Niger-Congo	Africa	14078,521
Indonesian	ind	id	Malayo-Sumbawan	Austronesian	Papunesia	19,972,325,222
isiXhosa	xho	xh	Bantoid	Niger-Congo	Africa	14,304,074
isiZulu	zul	zu	Bantoid	Niger-Congo	Africa	8,511,561
Kannada	kan	kn	Southern Dravidian	Dravidian	Eurasia	2,098,453,560

https://arxiv.org/pdf/2211.05100.pdf

Language	ISO-639-3	catalog-ref	Genus	Family	Macroarea	Size in Bytes
Kinyarwanda	kin	rw	Bantoid	Niger-Congo	Africa	40,428,299
Kirundi	run	rn	Bantoid	Niger-Congo	Africa	3,272,550
Lingala	lin	ln	Bantoid	Niger-Congo	Africa	1,650,804
Luganda	lug	lg	Bantoid	Niger-Congo	Africa	4,568,367
Malayalam	mal	ml	Southern Dravidian	Dravidian	Eurasia	3,662,571,498
Marathi	mar	mr	Indic	Indo-European	Eurasia	1,775,483,122
Nepali	nep	ne	Indic	Indo-European	Eurasia	2,551,307,393
Northern Sotho	nso	nso	Bantoid	Niger-Congo	Africa	1,764,506
Odia	ori	or	Indic	Indo-European	Eurasia	1,157,100,133
Portuguese	por	pt	Romance	Indo-European	Eurasia	79,277,543,375
Punjabi	pan	pa	Indic	Indo-European	Eurasia	1,572,109,752
Sesotho	sot	st	Bantoid	Niger-Congo	Africa	751,034
Setswana	tsn	tn	Bantoid	Niger-Congo	Africa	1,502,200
Simplified Chinese	—	zhs	Chinese	Sino-Tibetan	Eurasia	261,019,433,892
Spanish	spa	es	Romance	Indo-European	Eurasia	175,098,365,045
Swahili	swh	sw	Bantoid	Niger-Congo	Africa	236,482,543
Tamil	tam	ta	Southern Dravidian	Dravidian	Eurasia	7,989,206,220
Telugu	tel	te	South-Central Dravidian	Dravidian	Eurasia	2993407,159
Traditional Chinese	—	zht	Chinese	Sino-Tibetan	Eurasia	762,489,150
Twi	twi	tw	Kwa	Niger-Congo	Africa	1,265,041

https://arxiv.org/pdf/2211.05100.pdf

Language	ISO-639-3	catalog-ref	Genus	Family	Macroarea	Size in Bytes
Urdu	urd	ur	Indic	Indo-European	Eurasia	2,781,329,959
Vietnamese	vie	vi	Viet-Muong	Austro-Asiatic	Eurasia	43,709,279,959
Wolof	wol	wo	Wolof	Niger-Congo	Africa	3,606,973
Xitsonga	tso	ts	Bantoid	Niger-Congo	Africa	707,634
Yoruba	yor	yo	Defoid	Niger-Congo	Africa	89,695,835
Programming Languages	—	—	—	—		174,700,245,772

https://arxiv.org/pdf/2211.05100.pdf

Data Sources

Language Choices

- started with 8 languages, expanded Swahili, Hindi and Urdu;
- groups of 3 fluent in an additional language could add it

Source Selection

- “BigScience Catalogue”
- Arabic-focused Masader repositor
- 252 sources with at least 21 sources per language category
- Pseudocrawl for Spanish, Chinese, French, and English

GitHub Code (Google’s BigQuery)

OSCAR (38% of the corpus)

https://arxiv.org/pdf/2211.05100.pdf

Model Architecture

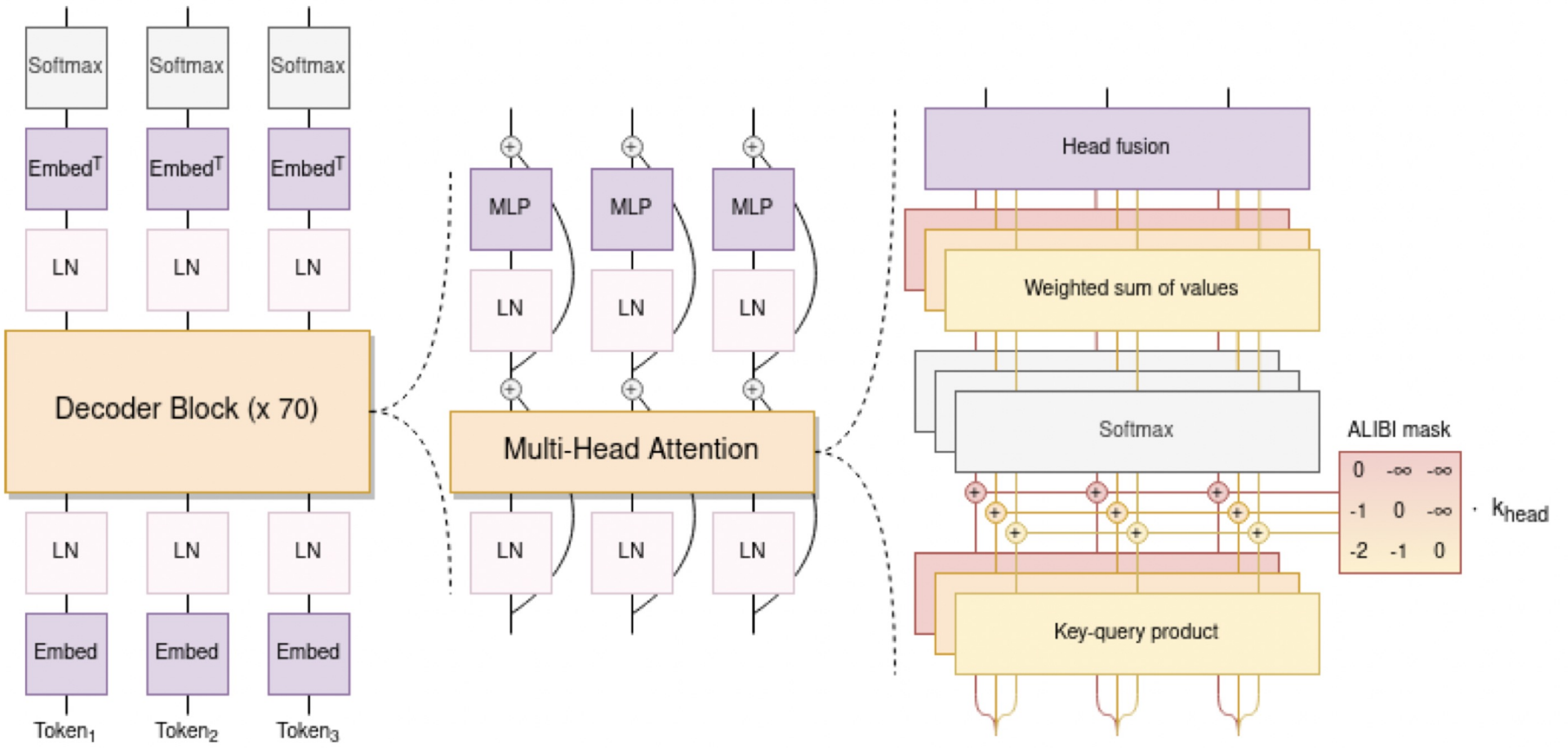


Figure 5: The BLOOM architecture. The k_{head} slope parameters for ALIBI are taken as $2^{\frac{-8i}{n}}$ with n the number of heads and $i \in 1, 2, ..., n$.

https://arxiv.org/pdf/2211.05100.pdf

Modeling Details

- ALiBi Positional Embeddings
- Embedding LayerNorm
- Vocabulary Size (250K)
- Byte-level BPE
- Pre-tokenizer (RegExp)

Tokenizer	fr	en	es	zh	hi	ar
Monolingual	1.30	1.15	1.12	1.50	1.07	1.16
BLOOM	1.17 (-11%)	1.15 (+0%)	1.16 (+3%)	1.58 (+5%)	1.18 (+9%)	1.34 (+13%)

Table 2: Fertilities obtained on Universal Dependencies treebanks on languages with existing monolingual tokenizers. The monolingual tokenizers we used were the ones from CamemBERT (Martin et al., 2020), GPT-2 (Radford et al., 2019), `DeepESP/gpt2-spanish`, `bert-base-chinese`, `monsoon-nlp/hindi-bert` and Arabic BERT (Safaya et al., 2020), all available on the HuggingFace Hub.

https://arxiv.org/pdf/2211.05100.pdf

Engineering

- trained on Jean Zay, a French government-funded supercomputer owned by GENCI and operated at IDRIS;
- 3.5 months to complete and consumed 1,082,990 compute hours;
- conducted on 48 nodes, each having 8 NVIDIA A100 80GB GPUs (a total of 384 GPUs);
- a reserve of 4 spare nodes with 2x AMD EPYC 7543 32-Core CPUs and 512 GB of RAM;
- trained on Megatron-DeepSpeed framework;
- bfloat16 mixed precision, which proved to solve the instability problem.

https://arxiv.org/pdf/2211.05100.pdf

Engineering

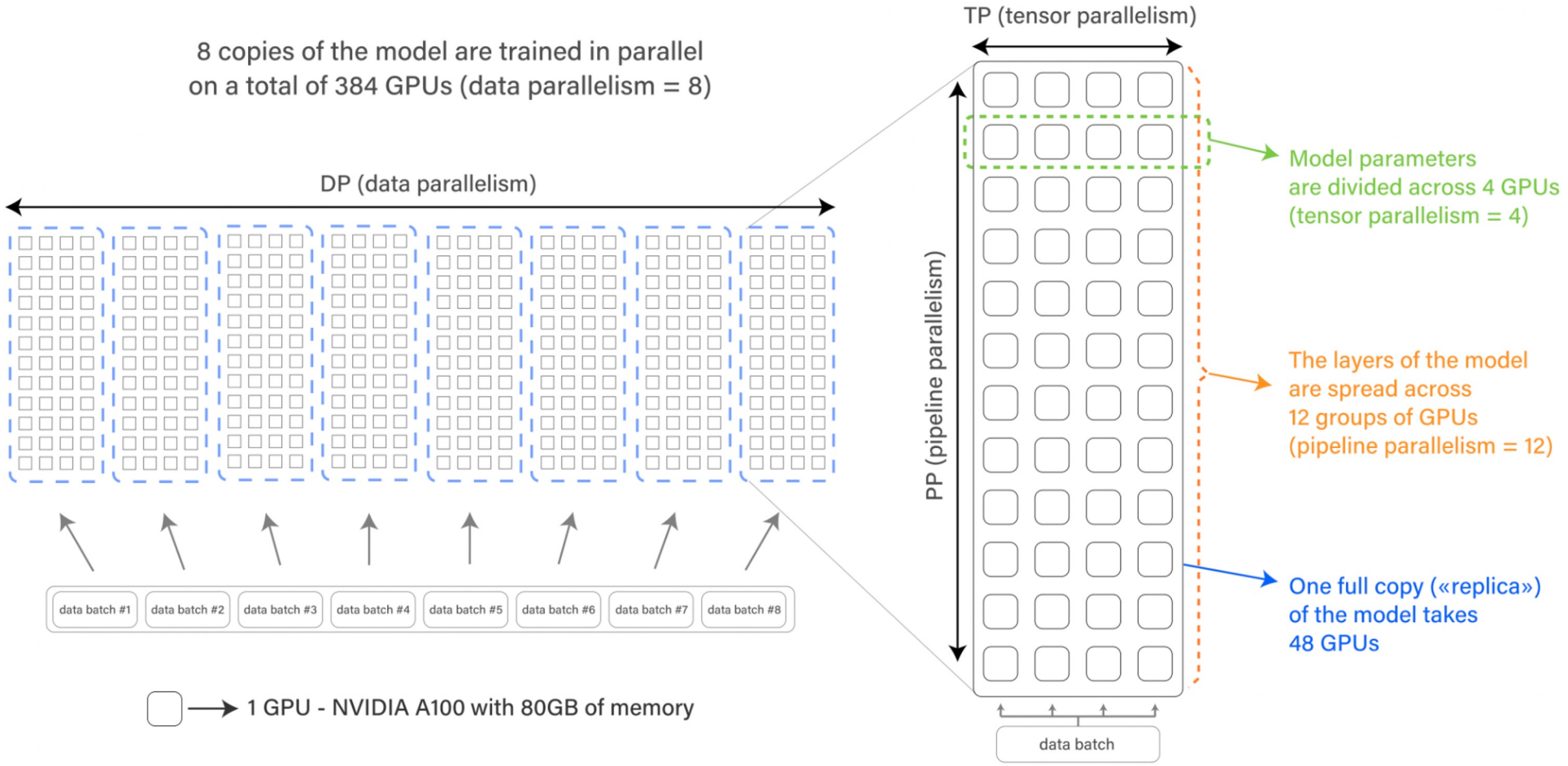


Figure 6: DP+PP+TP combination leads to 3D parallelism.

https://arxiv.org/pdf/2211.05100.pdf

SuperGLUE

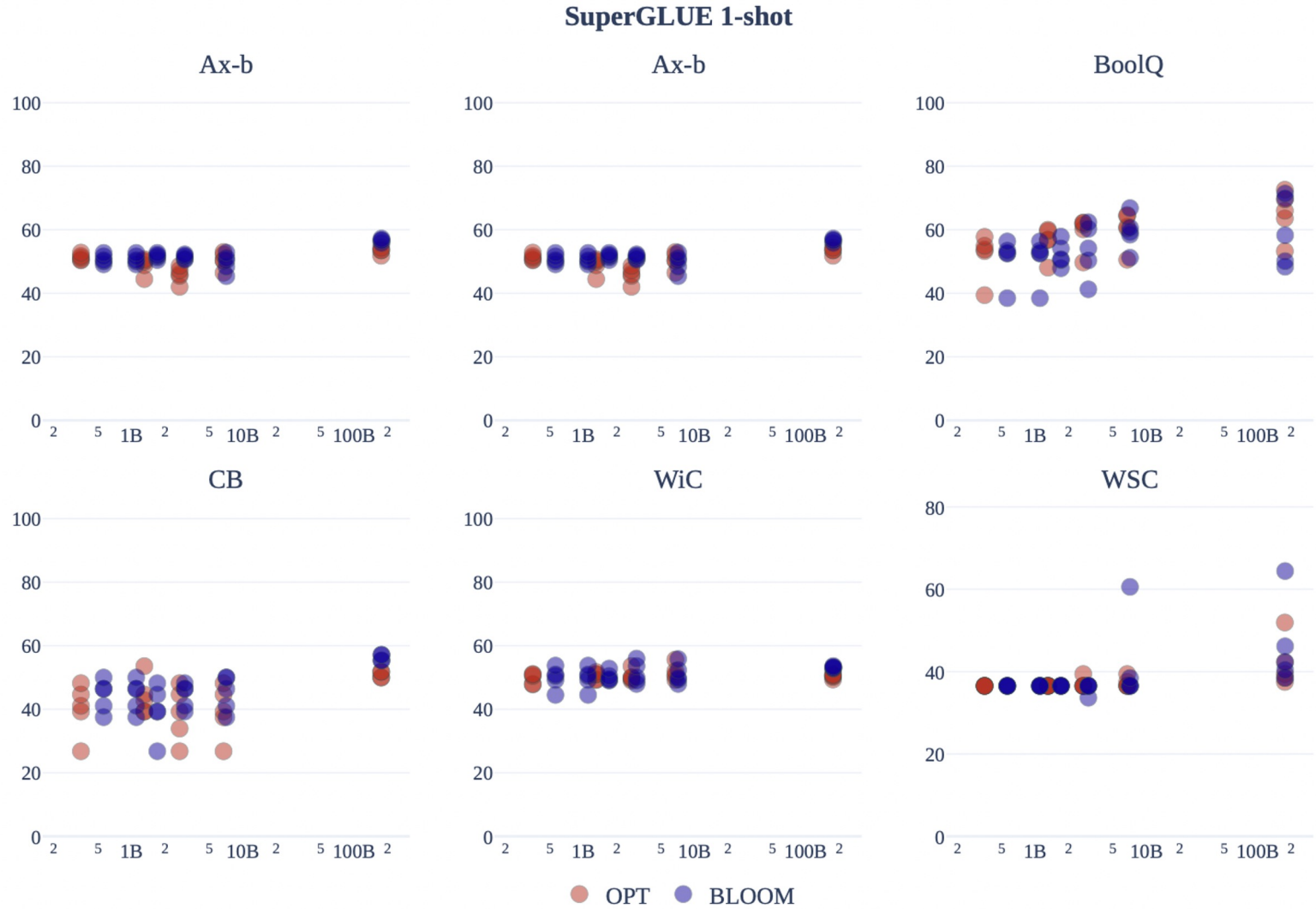


Figure 8: Comparison of the scaling of BLOOM versus OPT on each SuperGLUE one-shot task. Each point represents the average accuracy of a model within the BLOOM or OPT family of models on one of the five task prompts. The number of parameters on the x-axis is presented in log-scale.

https://arxiv.org/pdf/2211.05100.pdf

Machine Translation (1-shot)

Src↓	Trg→	eng	ben	hin	swh	yor
eng	M2M	–	23.04	28.15	29.65	2.17
	BLOOM	–	25.52	27.57	21.7	2.8
ben	M2M	22.86	–	21.76	14.88	0.54
	BLOOM	30.23	–	16.4	–	–
hin	M2M	27.89	21.77	–	16.8	0.61
	BLOOM	35.40	23.0	–	–	–
swh	M2M	30.43	16.43	19.19	–	1.29
	BLOOOM	37.9	–	–	–	1.43
yor	M2M	4.18	1.27	1.94	1.93	–
	BLOOM	3.8	–	–	0.84	–

(a) Low-resource languages

Src↓	Trg→	cat	spa	fre	por
cat	M2M	–	25.17	35.08	35.15
	BLOOM	–	29.12	34.89	36.11
spa	M2M	23.12	–	29.33	28.1
	BLOOM	31.82	–	24.48	28.0
glg	M2M	30.07	27.65	37.06	34.81
	BLOOM	38.21	27.24	36.21	34.59
fre	M2M	28.74	25.6	–	37.84
	BLOOM	38.13	27.40	–	39.60
por	M2M	30.68	25.88	40.17	–
	BLOOM	40.02	28.1	40.55	–

(b) Romance languages

https://arxiv.org/pdf/2211.05100.pdf

Machine Translation (1-shot)

Src ↓	Trg →	eng	fre	hin	ind	vie
eng	M2M	–	41.99	28.15	37.26	35.1
	BLOOM	–	44.4	27.57	38.75	28.83
fre	M2M	37.17	–	22.91	29.14	30.26
	BLOOM	45.11	–	17.04	29.50	31.66
hin	M2M	27.89	25.88	–	21.03	23.85
	BLOOM	35.40	27.83	–	–	–
ind	M2M	33.74	30.81	22.18	–	31.4
	BLOOM	44.59	29.75	–	–	–
vie	M2M	29.51	28.52	20.35	27.1	–
	BLOOM	38.77	28.57	–	–	–

(d) High→mid-resource language pairs.

Src ↓	Trg →	ara	fre	eng	chi	spa
ara	M2M	–	25.7	25.5	13.1	16.74
	XGLM	–	17.9	27.7	–	–
	AlexaTM	–	35.5	41.8	–	23.2
	BLOOM	–	33.26	40.59	18.88	23.33
fre	M2M	15.4	–	37.2	17.61	25.6
	XGLM	5.9	–	40.4	–	–
	AlexaTM	24.7	–	47.1	–	26.3
	BLOOM	23.30	–	45.11	22.8	27.4
eng	M2M	17.9	42.0	–	19.33	25.6
	XGLM	11.5	36.0	–	–	–
	AlexaTM	32.0	50.7	–	–	31.0
	BLOOM	28.54	44.4	–	27.29	30.1
chi	M2M	11.55	24.32	20.91	–	15.92
	XGLM	–	–	–	–	–
	AlexaTM	–	–	–	–	–
	BLOOM	15.58	25.9	30.60	–	20.78
spa	M2M	12.1	29.3	25.1	14.86	–
	XGLM	–	–	–	–	–
	AlexaTM	20.8	33.4	34.6	??	–
	BLOOM	18.69	24.48	33.63	20.06	–

(c) High-resource language pairs.

https://arxiv.org/pdf/2211.05100.pdf

Summarization (1-shot)

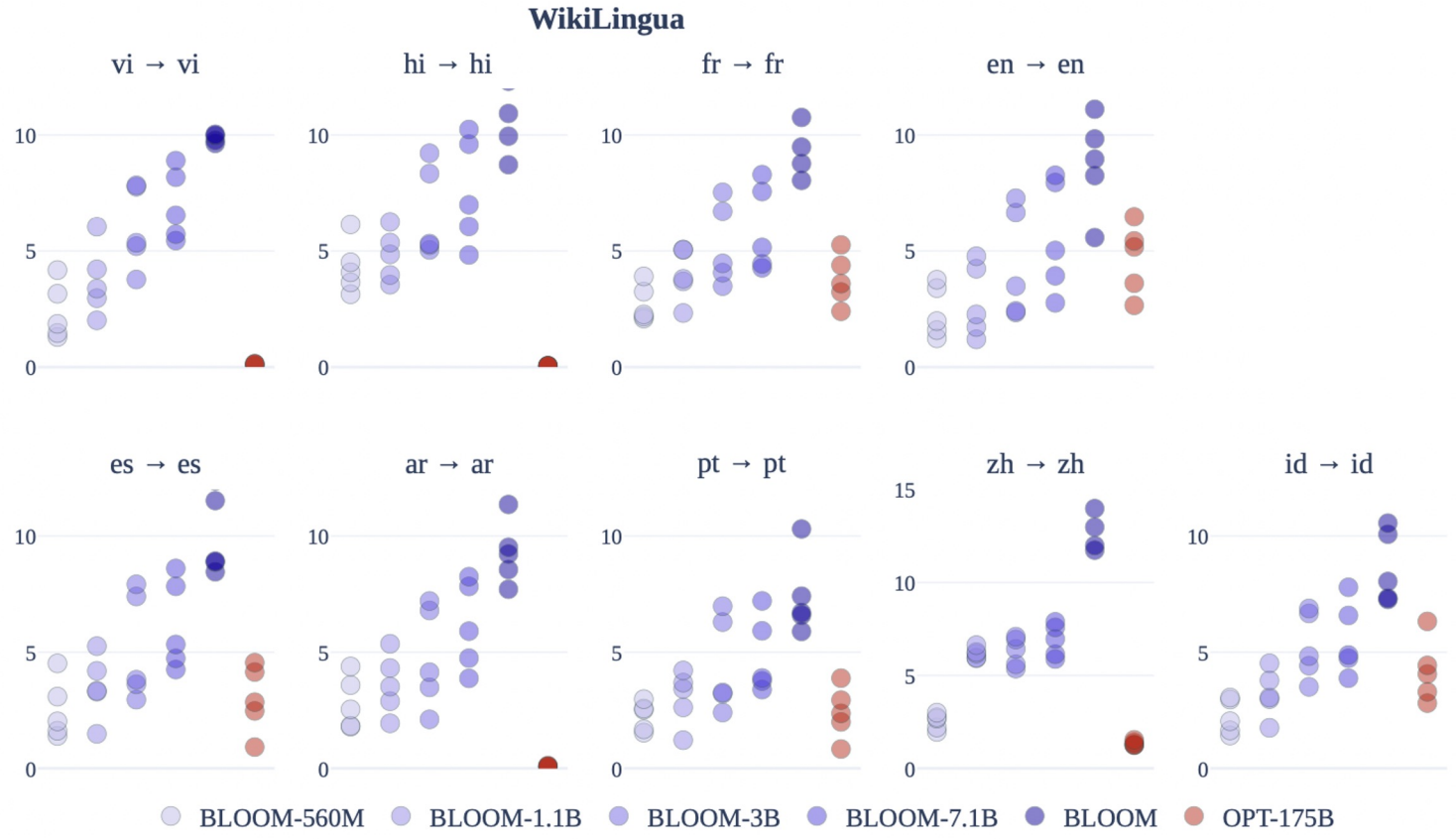


Figure 9: WikiLingua One-shot Results. Each plot represents a different language with per-prompt ROUGE-2 F-measure scores.

https://arxiv.org/pdf/2211.05100.pdf

BLOOM+1

- add more languages via
  - Continued Pretraining
  - MAD-X adapter
  - (IA)3 technique

Language	Language Family	Word Order	Script	Space-Separated	Seen Script
German	Indo-European (Germanic)	SVO	Latin	✓	✓
Bulgarian	Indo-European (Slavic)	SVO	Cyrillic	✓	✗
Russian	Indo-European (Slavic)	SVO	Cyrillic	✓	✗
Greek	Indo-European (Hellenic)	SVO	Greek	✓	✗
Turkish	Turkic	SOV	Latin	✓	✓
Korean	Koreanic	SOV	Hangul	✓	✗
Thai	Tai–Kadai	SVO	Thai	✗	✗
Guarani	Tupian	SVO	Latin	✓	✓

Table 1: Information about the unseen languages used in our experiments.

- we need around 100 million tokens of the new language for effective language adaptation
- when model sizes increases beyond 3 billion parameters, adapter-based language adaptation methods outperform continued pretraining

https://arxiv.org/pdf/2212.09535v1.pdf

Don’t have 8 A100s to play with?

An inference API, currently backed by Google’s TPU cloud and a FLAX version of the model, also allows quick tests, prototyping, and lower-scale use. You can already play with it on the Hugging Face Hub.

Conclusions

- Multilingual NLP is difficult and important
- Multilingual sentence encoders are an important resource
- There are multilingual encoder, decoder, and enc+dec transformers
- NLP resources for new languages can be bootstrapped